

METHODOLOGY AND SAMPLING ISSUES

FOR KPC SURVEYS

November 30, 1999

Eric Sarriot, MD, MPH, DTM&H

Peter Winch, MD, MPH

William M. Weiss, M.A.

Jennifer Wagman, MHS

Johns Hopkins University, School of Public Health, Department of International Health

Methodology and Sampling Issues for KPC Surveys is a publication of The Child Survival Technical Support (CSTS) Project. CSTS is funded by the Office of Private and Voluntary Cooperation, Bureau of Humanitarian Response (BHR/PVC), of the United States Agency for International Development (USAID), under contract number FAO-C-00-98-00079-00, and is managed by ORC Macro. The opinions expressed herein are those of the author(s) and do not necessarily reflect the views of USAID.

For further information on the Child Survival Technical Support Project please contact ORC Macro, CSTS Project, 11785 Beltsville Drive, Calverton, Maryland 20705; (301) 572-0823; e-mail: csts@macroint.com; Internet: www.childsurvival.com.

Table of contents

Introduction 4

SECTION 1 : Approaches to monitoring for results 6

1. Monitoring for results through population-based surveys 6
2. Development of the KPC survey 7

SECTION 2: Evaluation objectives of KPC surveys 10

1. Assessing progress towards objectives 13
2. Demonstrating change 16
3. Demonstrating causality 19
4. Improving program management 22

SECTION 3: Sampling options for KPC surveys 24

1. Cluster sampling 24
 - 1.1. Presentation 24
 - 1.2. What is the design effect? 28
 - 1.3. Decreasing the homogeneity within clusters 31
 - 1.4. Options for managing sampling questions within sub-groups 32
2. Management information at the local level and Lot Quality Assurance

Sampling (LQAS) 34

- 2.1. A comparison of LQAS and cluster-sampling 34
- 2.2. A brief review of LQAS applications in the literature 37
- Management information at the peripheral level 38
- Regional level coverage estimates by aggregating local data 38
3. Qualitative questions in the KPC survey 41
4. Annotated sampling examples 41

Conclusion 58

Resources on methodology and sampling issues for surveys 61

References 70

Abbreviations used

CI Confidence Interval

CS	Child Survival
DEf	Design Effect
DHS	Demographic and Health Survey
DPI	Detailed Program Implementation
EPI	Expanded Program of Immunization
IMCI	Integrated Management of Childhood Illnesses
KPC	Knowledge Practice Coverage
LQAS	Lot Quality Assurance Sampling
MICS	Multi Indicator Cluster Survey
MOH	Ministry of Health
PLA	Participatory Learning and Action
PVO	Private Volunteer Organization

Introduction

This report is intended for PVO child survival program managers as a methodological complement to the revised KPC survey. It is not a manual or a training guide for the implementation of the KPC survey, although it may be used in conjunction with other training materials. Its purpose is to clarify some of the methodological questions that have developed from the use of the KPC survey over the last 10 years. It consists of three sections:

Section 1: Approaches to program monitoring and evaluating for results with surveys

Objectives: *At the end of section 1, the reader will be able to:*
Distinguish three different types of results assessed through surveys.
Describe the initial and fundamental purpose of the KPC survey.

Section 2: Evaluation objectives of KPC surveys

Objectives: *At the end of section 2, the reader will be able to:*
Present three different levels of evaluation for results using KPC survey data.
Offer methodologically-appropriate strategies for conducting each level of results evaluation.
Discuss the implications and the feasibility of each evaluation design from a program management perspective.

Section 3: Sampling options for KPC surveys

Objectives: *At the end of section 3, the reader will be able to:*

Explain how the cluster-sampling method was adapted for the KPC survey, from the EPI 30-cluster sampling method.

Describe what the design effect (Def) in cluster sampling is, and its importance for program evaluation.

Choose an appropriate strategy to solve common methodological problems encountered with cluster sampling for KPC surveys.

Have a clear understanding of Lot Quality Assurance Sampling (LQAS), its basic principles, and its application in the assessment of child survival programs.

Compare the respective strengths and weaknesses of cluster vs. lot quality assurance sampling.

Finally, some guidelines and manuals on survey and sampling issues are presented to the reader in the annex of the document.

The structure of this report is summarized in figure 1.

Figure 1: structure of the report:

Approaches to monitoring for results
Evaluation objectives of KPC surveys
Sampling options for KPC surveys
Conclusion
Resources on methodology and sampling issues for surveys

SECTION 1 : Approaches to monitoring for results

Evaluation objectives of KPC surveys

Sampling options for KPC surveys

Conclusion

Resources on methodology and sampling issues for surveys

Objectives: *At the end of section 1, the reader will be able to:*
distinguish three different types of results assessed through surveys;
describe the initial and fundamental purpose of the KPC survey.

1. Monitoring for results through population-based surveys

The achievements of child survival programs can be categorized as:
activities (inputs and processes) conducted by the program,
objectives: from outputs (benefits directly and immediately received by the population of intervention) and outcomes (intermediate results) to final results (long term impact),
Goals such as reduction in mortality are beyond the responsibility of one single project.

Program or service monitoring systems will measure activities. Immediate, intermediate and long term objectives or results are usually assessed by surveys of the communities where the intervention took place. Health impact can be measured by anthropometric surveys, and morbidity and mortality surveys, such as the DHS survey (figure 2). The Knowledge-Practices-Coverage (KPC) survey, and others such as UNICEF's Multiple Indicator Cluster Survey (MICS) survey, focus on program outcomes such as the knowledge and practices promoted in the 'target community', as well as on indicators of coverage for the services delivered (e.g. immunization coverage). Each level of measurement requires a specific tool, for which an appropriate sampling scheme has to be designed. The focus of this report is on monitoring outcomes and results through population-based surveys, specifically KPC surveys. Its development, over 10 years ago,

and its fundamental purposes will now be discussed.

Figure 2: Surveys for monitoring for results

2. Development of the KPC survey

The KPC survey was developed 10 years ago, at the request of USAID, in order to provide a uniform approach to baseline and final data collection in child survival programs of PVOs working in different countries {PVO Child Survival Technical Report; 1993}. Before its development, the prevailing situation was that few projects had conducted any type of survey by the time of submission of their DIP (Detailed Implementation Plan). Those which did used samples of varying sizes, measured different variables and had different purposes for the survey they conducted. The KPC survey became a requirement for all baseline and final assessments for some years after 1991. (Even though it is still widely used, it is no longer a requirement for USAID-funded projects). Its purpose was, from the start, to be a standardized, scientifically valid and reliable, low-cost management and evaluation tool. It had to allow quick data collection by field staff (in about 20 minutes per household) on key behavioral and coverage indicators, susceptible to change over the period of a project.

Because it is impossible to survey an entire population, survey evaluation methods have to rely on extracting a sample from this population to conduct the analysis. We rely on chance (randomization) to *limit* the bias in the selection of the sample. If all members

of the population have the same chance of being drawn into the sample, we expect that our sample will be representative of the entire population. This expectation needs to be confirmed by a comparison of the sample's characteristics (demographic, socio-economic, urban vs. rural, etc.) with the information available about the larger population.

Because we rely on a sample, we will only be able to estimate the true percentages achieved for each indicator. Statistical principles teach us that the greater our sample size, the greater the precision of our estimate. Specific formulas are available in all statistical manuals, and guide the evaluator in the selection of a sample size allowing to calculate an estimate or to make a comparison with a theoretical value (for example a performance objective).

Cluster sampling was proposed as a reliable and cost-efficient way to gather the information needed, and has been the primary sampling method used in KPC surveys over the last 10 years. This sampling method was selected assuming that the data collected would be used for the purposes of decision-making and program management. The KPC survey was never expected to be a tool to address research issues or gather in-depth social and demographic data, which would require different questions and different sampling approaches.

During the last 10 years KPC cluster-surveys have considerably improved the ability of PVO Child Survival projects to identify priorities, define objectives based on data, and measure progress towards these objectives. KPC cluster-surveys were never expected, much less designed to measure change between two periods of time, or to compare different groups of population in order to demonstrate that a specific intervention was the cause of an observed change. (figure 3). Assessing change over time and demonstrating causality have very specific requirements and constraints, which will be discussed in section 2.

Figure 3: Realistic and unrealistic expectations about KPC surveys (under the standard 30-cluster design, with 10 observation units per cluster)

() The KPC survey does not forbid attempts to make such demonstrations, providing modifications of its design, but it was not conceived in its standard use with 30 clusters of 10 households to be able to answer these questions satisfactorily for all indicators.*

SECTION 2: Evaluation objectives of KPC surveys

Sampling options for KPC surveys

Conclusion

Resources on methodology and sampling issues for surveys

Objectives: *At the end of section 2, the reader will be able to:*

Present three different levels of evaluation for results using KPC survey data.

Offer methodologically-appropriate strategies for conducting each level of results evaluation.

Discuss the implications and the feasibility of each evaluation design from a program management perspective.

We will now consider the three phases of a project at which a KPC survey can usually be implemented and then describe three levels of program evaluation, or three types of evaluation questions that program managers and evaluators may wish to address:

Have the program objectives been reached?

Has change taken place in the intervention population for selected indicators?

Did the program cause the observed change?

It is essential to emphasize at this point the different perspectives of statisticians and managers about these questions. While the former focus on rejecting hypotheses based on probabilistic approaches, the latter are driven by a need to extract meaning from the survey data in order to come to a decision. We will stress how these two perspectives affect methodological and sampling issues throughout our discussion.

A KPC survey can take place at three points in the life of a project: at baseline and at the end of the project, but also during the life of the project, somewhere around a mid-point.

Baseline survey

The baseline survey provides a general profile of the population of intervention with regards to key indicators. Uncertainty and subjectivity in the determination of priorities and objectives are reduced by the survey data.

Mid-term evaluation

Mid-term evaluations do not always include a KPC survey. They may rely on the service level monitoring data, and vary in the depth of their analysis. There is a broad range of questions that monitoring can try to answer, from specifying exactly what services are being delivered, to assessing the quality of the services delivered and, finally, measuring the results at the community level.

Obviously the level of complexity increases as one tries to answer each question. The data may be available from service monitoring records, direct observation and interviews of 'clients', or rapid assessments combining focused survey questions and qualitative research tools. Surveys will not always provide the most appropriate answers and it is unlikely that they will be conducted with samples large enough to allow any comparison with the baseline. Evaluators and managers should focus on the appropriateness and usefulness of the questions, rather than on issues of power and sample size. The impact of a project's IEC component might be constrained by important cultural issues, which will be explored more efficiently by observation for example, than by survey questions. A few structured 'live observations' of nutrition programs' graduates in their homes for example, might shed more light on the constraints faced by the program than any survey question possibly could.

Mid-term evaluation should then focus on providing a general sense of what the program's accomplishments are, and answering key qualitative questions about the services delivered.

Final Evaluation

Finally, the KPC survey can be conducted at the end of the intervention, as part of a final evaluation. Managers will try to assess the results of the program and possibly to answer different questions through this final (or summative) evaluation (figure 4):

Were objectives reached in the region of implementation of the program?

Can the program demonstrate an improvement over time in knowledge, practices, or coverage, from the baseline levels?

Can the program demonstrate its specific responsibility in an observed change between baseline and end-point evaluation?

Obviously, these questions make sense only if the program *activities* have been conducted as planned. Monitoring records, process evaluation, and mid-term evaluation data, if they are available, will provide information about the basic question of the delivery of services. Our discussion focuses here on the *achievements* of the program and the related evaluation questions.

Ideally, program managers would like to be able to answer all these questions for themselves and for the donor agency. But this is not what the KPC cluster-survey was designed to do (in its standard format of 30 clusters of 10 households each). It is important to fully understand that, while the baseline data allow the manager to set reasonable objectives for the program, the 30-cluster design does not establish the most appropriate baseline for comparison point with data from the final KPC survey. Comparing a final estimate to a fixed objective, comparing this estimate to a previous estimate and establishing the causality of the program in a demonstrated change are three very different questions, constrained by different methodological requirements, which will now be discussed.

Figure 4: Levels of program evaluation and increasing methodological constraints

1. Assessing progress towards objectives

A sample size of 30 clusters of 10 households is sufficient to provide an estimate with $\pm 10\%$ precision for a coverage rate, and can establish with reasonable confidence (determined by the α -value traditionally set at 5%) whether an objective has been met or not. Because the estimate is compared with a fixed value (the program performance objective), only one level of imprecision has to be taken into account (figure 5).

Let us consider that, based on our sample of 300 survey respondents, we have obtained an estimate of 80% for the coverage rate on this indicator. The meaning of choosing a confidence level of 5%, and achieving a precision level of $\pm 10\%$ is that:

The true coverage rate in the entire population (which we do not know but estimate at 80%) has a 95% chance of being within 10% of our estimate, (in this case between 70% and 90%).

Let us now consider four possible levels of objectives that our project could have initially set out to reach:

- Objective A (e.g. 68%) is outside and below the confidence interval of our estimate,
- Objective B (e.g. 75%) is below our estimate but within its confidence interval,
- Objective C (e.g. 85%) is above our estimate but within its confidence interval,
- Objective D (e.g. 93%) is outside and above the confidence interval of our estimate.

Table 1 suggests the conclusions that can be made, from a statistical perspective and from a management perspective about these four different situations.

Figure 5: Comparison of a final estimate to a pre-set objective (hypothesized at four different values)

Table 1: Assessing achievement of objectives

Level of objective	Statistical Conclusion	Managerial Conclusion
Objective A	The difference between our best estimate and objective A is statistically significant. We reject the hypothesis that the population coverage rate is equal to the pre-set objective. We are more than 95% confident to have reached our objective.	Our objective has been reached. → Continue activities, or → Expand program, or → Plan transfer and phase-out → etc.
Objective B	We cannot reject the hypothesis that the population coverage rate is equal to the pre-set objective. Our best estimate is that the population coverage rate is 5% higher than objective B, but the difference is not statistically significant.	Our objective has probably been achieved (our best estimate is that we are 5% above objective). There is no evidence that we have failed to reach our objective, but we cannot prove that we have achieved coverage higher than the objective. → Was performance homogenous in all local areas of interventions? → Do other sources of information support or contradict our conclusion?
Objective C	We cannot reject the hypothesis that the population coverage rate is equal to the pre-set objective. Our best estimate is that the population coverage rate is 5% lower than objective C, but the difference is not statistically significant.	Our objective has probably not been achieved (our best estimate is that we are 5% below objective). There is no statistical evidence that we have failed to reach our objective, but we cannot prove that we have achieved coverage higher than the objective. → Was the objective too ambitious? → Was performance homogenous in all local areas of interventions? → Do other sources of information support or contradict our conclusion?
Objective D	The difference between our best estimate and objective D is statistically significant. We reject the hypothesis that the population coverage rate is equal to the pre-set objective. We are more than 95% confident that our program has not reached its objective.	Our objective has not been reached. → Was the objective too ambitious? → Was low performance homogenous in all local areas of interventions? or → Are specific local areas responsible for overall low performance? → Plan and implement corrective measures

As illustrated in this theoretical example, once a hypothesis is formulated (e.g. an estimate is equal to or superior-or-equal to a set value), statistical tests achieve a high level of certainty when they can disprove this hypothesis. On the other end, managers look for positive evidence that they have achieved this objective. This ambiguity cannot be totally resolved, but a better understanding of its nature will help managers make better use of survey results.

Defining objectives and assessing achievement of objectives belong in the field of management and decision-making. Relatively small samples are allowed, and the data can be collected and computed by field-staff. The 30-cluster sampling scheme for the KPC survey was chosen to achieve this level of precision and reliably answer these management questions.

2. Demonstrating change

Comparing a final estimate to a set objective only introduces one level of imprecision: that of the estimate. But, in order to compare final and baseline estimates, two levels of imprecision, those of the baseline estimate and the final estimate, have to be taken into account, and require an increased sample size *for the two* surveys. (The reader is encouraged to refer to the appropriate statistical textbooks for a better exposition of the statistical principles involved. Annotated numerical examples of sample size calculation for comparison with a pre-set objective and comparison between two samples are provided in section 3.)

While in the previous example, we were testing whether a fixed value was below or within the confidence interval for our estimate, we are now testing whether two estimates, each with its own imprecision, are statistically different. In the first case, we established with 95% confidence that the objective of 68% coverage (objective A) was not in the confidence interval of our estimate (70% to 90%). In this second situation, we must calculate a confidence interval for the difference between the two samples. (We can also test whether this difference is significantly different from zero, the two methods being

equivalent.) In order for us to statistically demonstrate a difference between the two estimates, we will need a larger sample at each phase, thus reducing the two levels of imprecision.

Figure 6: comparing estimates from two sample

The decision to make a before-after comparison must be taken before the onset of the program, and the sample size of the baseline survey must be calculated appropriately. (Correcting the final survey sample size for a late decision to ‘power the survey’ in such a way as to be able to make a comparison is sometimes possible, but not entirely recommendable.) Although the desire of program managers to compare the final and baseline estimates for different KPC indicators is understandable, they should only do so under the following conditions:

First, establish whether the confidence interval for the final estimate excludes the program objective for a given indicator. Statistical evidence should be used when it exists.

This first level of assessment is important in establishing with what level of certainty the program is thought to have reached its objectives (refer to table 1).

Then, if a comparison of baseline and final results is presented, the evaluator should make explicit whether the sample size for the two surveys did or did not attempt to be able to demonstrate a difference with statistical significance.

An observed difference should be reported with its confidence interval (see annotated sampling example # 5).

Finally, inasmuch as possible, other sources of information should be used to try and assess whether the observed difference may or may not be genuine.

What type of demonstration (probable or plausible) can be made about the observed difference in estimates?

Managers and evaluators must be clear about the type of demonstration they are trying to make. Unless they set themselves, from the inception of the program on, to do so, they should *not* try to demonstrate change probabilistically. *Probabilistic* demonstration can refer, however, to the observed achievement of objectives in the populations of intervention when the samples’ size provide enough power for a

comparison test. In the case of the comparison of the two estimates, with 30-clusters of 10 households, given the size of the Design Effects (to be discussed further below) the sizes of the two samples are not expected to provide enough statistical power for a probabilistic demonstration of change. In other words, the observed change between baseline and final evaluation will usually not be statistically significant.

What managers and evaluators can attempt to support, however, is the *plausibility* of a change in the population, based on the spectrum of data available for their consideration.¹

What other sources of information can be used to *triangulate* with the data collected?

Triangulation means using different sources of information in order to support the evidence for an observed difference (or lack thereof). The evaluator may refer to other surveys on national trends, focused research in the program intervention area, trends observed in non-intervention areas with similar cultural, social and demographic characteristics, in order to support the plausibility of a purported change between the baseline and final situations.

The complexity increases when there is a larger national and secular negative trend for a given indicator, which the program attempts to limit or correct, for example when immunization coverage at the national level is decreasing. In this situation, a lack of difference between the two phases would be a positive outcome, and it is by reference to other reliable data sources, or to similar measures in control areas, that a program can suggest its impact. Qualitative and ‘subjective’ analyses are going to be part of an evaluator’s report in such a case. Rigorous evaluation requires being explicit about the assumptions and limitations of the evaluation, however, and not making inappropriate use of statistical inference, when the study design does not allow for it.

3. Demonstrating causality

Demonstrating the causal role of an intervention in effecting an observed change requires that other (spurious) factors that could also have produced the change be accounted for in the study design. This is the level of analysis with which researchers are concerned, far more than managers. The results achieved in the areas of intervention must be compared with those in non-intervention or control areas. Causality will be attributed to the intervention if the following conditions are fulfilled:

A positive change has been observed in the regions of intervention.

Exposure or service coverage indicators are high in the regions of intervention.

Non-intervention regions have not reached the same levels of exposure, or service coverage.

The change has not occurred with the same amplitude in the non-intervention regions.

The intervention and non-intervention regions are comparable with regard to all other factors.

Spurious factors of change, and possible threats to validity have been accounted for.

These questions cannot be answered outside of strictly controlled study designs (see box 1) where the control areas are comparable to the intervention areas and allocation to the control or experimental group is determined at random (see box 2). This rarely corresponds with the mix of epidemiological, pragmatic, and political factors influencing the choice of a region of intervention in a PVO Child Survival project. Pseudo-experimental study designs (selection of a control area without randomization) are frequently the only alternative possible, but increase the level of uncertainty about the validity of the findings. Large sample sizes allowing comparison between time period and monitoring in different areas (both intervention and control) are also required.

In some situations, when randomization takes place at the community level and not at the individual level, it is recommended for evaluators to use statistical tests based on the level where randomization has taken place.² This issue cannot be treated in depth in this report, but a practical example is provided in the next section (example 6 in section 3).

There are many potential threats to the validity of the inferences made that the evaluator will also need to consider (box 3). It is not the purpose of this paper to discuss them now, but only to emphasize that demonstrating causality requires a high investment in measurement and analysis, much beyond what is expected of child survival program evaluation.

Box 1: Why are control groups needed?

Human behaviors and health status indicators are not elements fixed in time. Many factors will influence them, from national policies, to local events and processes, not to forget the effects of mass media, migration, new infrastructures in the local economy, etc. It would not be unreasonable to assume that an indicator measured through a population-based survey would remain unchanged over the course of three to five years. Practice and coverage indicators may increase or decrease during the usual period of time between a baseline and a final evaluation, independently of our programs.

Control groups allow us to compare the results in the intervention areas with those due to all these other factors, as they are observed in the control areas. The specific impact of a child survival intervention will be given by a comparison between control and intervention zones. If a given indicator has failed to improve throughout the intervention, the program may still benefit from a positive evaluation if the trend has been on a decrease in coverage for this particular indicator in the control zones. Inversely, there would be no reason to credit a program for a measured increase in a breastfeeding practices in the intervention zones, if this increase has a similar amplitude in control areas (presumably due to another and larger program, or to a secular trend).

Box 2: Why use random procedures to select groups?

The entire premise of a comparison between control and intervention areas rests on the comparability of the two types of areas. Control areas should be similar to intervention areas in all things, except for benefiting from the intervention. Because each area has the same chance of being selected for the intervention as any other, random selection is our best tool to reduce all the possible biases affecting a comparison.

Inversely, if a region is chosen for an intervention because of political factors, or because of the pre-existence of partners and networks, it is likely (or at least it cannot be reasonably excluded) that the same factors that put the region in a position to be 'favored' would also prevent it from being compared to non-intervention areas.

Box 3: What are threats to validity?

Even when all the conditions for an experimental design have been respected, and a specific indicator is improved in the intervention areas statistically significantly more than in non-intervention areas, factors external to the study can interfere with the observed difference in change. These factors are called threats to validity, as they may render the study's conclusions about the effect of the intervention invalid. These threats to validity may be time-related change in the indicator among either of the study groups, instrumentation, sensitization, and other method-related biases, population changes in either type of group, 'contamination' of the control areas by elements of the intervention, etc. Only careful and systematic consideration of each possible threat will help the evaluator estimate or rule out their influence, and support the validity of the evaluation findings.

Figure 7 summarizes the differences and increasing complexity of the different levels of analysis in monitoring program results.

Figure 7: Levels of program evaluation and increasing methodological constraints:

4. Improving program management

The appropriate level of (im)precision for management needs

Evaluation methods need to be tailored to the decision making process. Because many factors will remain uncontrollable in the evaluation design, managers and evaluators must start from the endpoint of the decision and then systematically consider:

- which questions need to be answered,
- how survey or other methods can provide this information,
- what level of precision in the measurement is to be achieved with the resources available,
- what other sources of information can shed light on the qualitative process of sifting through the data to assess the results achieved by the intervention.

Imprecision cannot be eliminated from evaluation, but programs should make reasonable decisions about how much precision is needed. A larger sample size will increase the precision of an estimate. But this increased precision will come at a certain cost. If this cost is too high, evaluators and managers should consider alternative strategies for collecting information and should not expect a 30-cluster survey to be the answer to all their needs.

Evaluation as a capacity-building exercise

Local capacity-building is a key purpose for the development of better monitoring and evaluation systems in CS projects. Developing the local capacity to base decisions on reliable data, through a critical review of all the information available, is to be a goal of every program. The sustainability of the information systems being designed is consequently one central question to be addressed in selecting an evaluation method.

By involving local partners in the design or adaptation of the KPC survey as well as in the survey implementation and data analysis, KPC surveys provide a way to raise the awareness of local partners on the critical problems targeted by the program. Any 'improvement' on the evaluation design should be weighed, not only in terms of the added information benefits, but also by considering the increased complexity of data collection, computation and analysis. Part of the responsibility of managers in program assessment is to ensure that the evaluation systems being developed will be transferable to national counterparts along with the rest of program management responsibilities.

Approaches to monitoring for results

Evaluation objectives of KPC surveys

SECTION 3: Sampling options for KPC surveys

Conclusion

Resources on methodology and sampling issues for surveys

Objectives: *At the end of section 3, the reader will be able to:*

Explain how the cluster-sampling method was adapted for the KPC survey, from the EPI 30-cluster sampling method.

Describe what the design effect (DEf) in cluster sampling is, and its importance for program evaluation.

Choose an appropriate strategy to solve common methodological problems encountered with cluster sampling for KPC surveys.

Have a clear understanding of Lot Quality Assurance Sampling (LQAS), its basic principles, and its application in the assessment of child survival programs.

Compare the respective strengths and weaknesses of cluster vs. lot quality assurance sampling.

1. Cluster sampling

1.1. Presentation

The most commonly used probability sampling methods are:

Simple random sampling,

Stratified random sampling,

Systematic sampling, and

Multistage cluster sampling.

A simple random sample is the most direct way to extract a sample from the population. The sampling frame must include each and every individual in the population. This requires a comprehensive demographic data base or inventory of households and, in the setting of a developing country, would induce very large expenses to survey a very small number of persons in any one village or district.

A systematic random sample is very similar to a simple random sample, but differs from it by the method of extraction of the individuals from the sampling frame.

In stratified random sampling, the population is divided into strata, or sub-groups, and a random sample is extracted from each sub-group. The definition of the sub-groups depends entirely on the theoretical basis of the intervention, and its expected impact on the different strata defined. Stratified random sampling brings an additional level of complexity and cost to the sampling process, but it ensures that all pertinent sub-groups

will be included in the sample.

Cluster sampling, which will be discussed in greater depth as it applies to KPC surveys, involves the random selection of clusters (naturally occurring units such as villages, city blocks, schools) and the selection of all members of the selected cluster in the sample.

In multistage cluster sampling, simple random sampling is used within the cluster to select the members of the sample.

Figure 8 provides a comparison of simple random sampling and cluster sampling.

Figure 8: a comparison of simple random sampling and cluster sampling

Multistage cluster sampling makes use of the existence of natural groups where populations aggregate: schools, hospitals, villages, etc. The first stage of sampling consists of randomly selecting a number of such clusters. In a second stage, individuals or households are randomly selected within the clusters. This approach allows the randomization process to take place on a sampling frame made of villages, urban districts, or whichever cluster unit has been defined, instead of a comprehensive list of the population of the region of interest.

WHO and UNICEF developed the EPI 30-cluster sample method to assess the immunization coverage at a national level in a cost-effective and rapid way.³ In this approach, after the selection of 30 clusters, seven households are randomly selected within each cluster, and available mothers of children in the 12-23 months age group are surveyed. This yields a total sample size of 210 children and provides an estimate within 10 percentage points of the population immunization rate. The determination of the number of clusters and of the total sample size for the EPI surveys, to achieve the $\pm 10\%$ level of precision, are explained in figure 99.

A simple random sample of size 96 would produce this level of precision, but require the visit of 96 different sites by the surveyors. By randomizing clusters at the first stage, the number of sites visited is limited to 30 clusters. This number of clusters was chosen as the minimal number necessary to respect the Central Limit Theorem, a key principle in statistics. More clusters would increase the precision and the cost of the survey, while less than 30 clusters would lead to too great a risk of misestimating the true coverage for the area. As will be presented in the next sub-section, the clustering of the data collected introduces an additional level of imprecision, which requires increasing the sample size in order to maintain the level of precision. The EPI sample size of 210 children is thus a compromise between various types of constraints: methodological and logistical.³

The KPC survey involves a larger age group of children as it targets children aged 0 to 24 months. To maintain an approximate $\pm 10\%$ level of precision for most variables (remembering that the sample size, level of precision and DEf are all variable-specific), the

minimum sample size for the KPC survey was set at 300 (30 clusters of 10 children) instead of 210 (30 clusters of seven children), yielding average sub-sample sizes around 120 for 12-23 months children, and 60 for the age-group 0-6months.

Important points to note at this stage are that:

- There is nothing particularly scientific in the choice of a $\pm 10\%$ margin of error. It was chosen as a seemingly reasonable compromise between the costs and advantages of gaining more precision.
- Each program should determine when a higher level of precision is required and increase the sample size accordingly (options for doing this are also discussed below).

Figure 9: Comparison of two ways of calculating sample sizes for immunization coverage estimates

1.2. What is the design effect?

Cluster sampling introduces the notion of a design effect (DEf). As individuals are more likely to share common characteristics within clusters than if selected by direct random sampling, a new level of imprecision is added in the calculation of overall estimates. This added level of imprecision is called the design effect. The DEf can be calculated *after* the study as the proportion of the sample variance due to clustering over the variance of a non-clustered random sample. (In some cases this ratio of the variances is referred to as DEf^2 and called the “design effect”, while the ratio of the standard errors is called the “design factor”. This distinction is not always clear in the literature.) It can also be estimated from previous studies on similar indicators, and with similar frequency distributions. Inversely, it is used to correct the sample size needed to achieve a given level of precision. In the absence of specific information about its value, it is usually assumed to be equal to two in EPI and KPC surveys. But in each specific setting and for each indicator, statistical software such as EPI-INFO's CSAMPLE, allows us to estimate the expected DEf for a final evaluation, from the results of the baseline survey. (The KPC Survey Trainer's Manual provides a formula to calculate the confidence interval of an estimate by hand by estimating the DEf (e.g. based on the results of the baseline evaluation). See appendix to this report.)

The DEf has certain characteristics:

It increases with the measurement of continuous data (e.g. weights or heights) as compared to dichotomous data (e.g. immunized Yes/No).

It decreases with stratification within clusters, because stratification reduces the homogeneity within the cluster.

It increases with the length of the data collection time.

It depends on the prevalence of the disease or indicator being measured.

It increases with the cluster size, and with differences in cluster sizes.

It increases with the magnitude of the disease association within cluster.

An in-depth discussion of DEf issues can be found in other publications.⁴⁻⁶ But it is essential to emphasize that:

Cluster sampling increases imprecision.

The design effect must be taken into consideration when building a confidence interval for an estimate.

The D_{ef} value can vary substantially depending on the indicator investigated and on the context⁷⁻¹¹, a study of D_{ef} observed on a number of KPC indicators in different CS projects finds that D_{ef} are closer to 1.5 than 2.0 for many variables (table 2). (Weiss W.M., 1999, unpublished)

The same study suggests that baseline and final surveys should be conducted during seasons of high diarrhea prevalence, in order to obtain sufficient precision on ORS and ORT practices.

Numeric examples of the use of D_{ef} in calculating a sample size are provided in section three.

Table 2: D_{ef} values encountered in different CS projects for specific indicators in CS projects in Nigeria, Indonesia, Honduras, Bangladesh, Papua New Guinea, and Honduras (Weiss W.M., 1999; unpublished)

Indicator	Range of D _{ef}
ORT use	1.06 - 1.79
Measles Immunization	1.06 - 2.16
Exclusive breastfeeding < 6mo	1.08 - 1.91
TT2	1.45 - 2.81

In considering the sampling options in a KPC survey, the evaluator can be faced with four additional types of methodological questions:

How can the precision of the estimates be increased in spite of the D_{ef}?

How can sub-groups in the sample (for example children 0-6 months old) be large enough to measure progress on a set of age-specific indicators?

Can the survey provide information about local districts, or determine which areas are reaching their objectives or not?

How to obtain reliable information about sensitive or complex questions unlikely to be answered through a survey question?

These four questions will guide the rest of our discussion and are presented in figure 10.

Figure 10: Options for improving on the 30-cluster KPC method

Question	Option (Section)
How can the precision of the estimates be increased in spite of the DEf?	Decrease the homogeneity within clusters. (1.3.)
How can sub-groups in the sample (for example children 0-6 months old) be large enough to measure progress on a set of age-specific indicators?	Strategies for ensuring a sufficient sub-group sample size (1.4.)
Can the survey provide information about local districts, or determine which areas are reaching their objectives or not?	Alternative sampling strategy: Lot Quality Assurance Sampling (2.)
How to obtain reliable information about sensitive or complex questions unlikely to be answered through a survey question?	Use of qualitative methods and non probability sampling.

1.3. Decreasing the homogeneity within clusters

The more respondents to the survey will be similar to one another with regards to the key indicators, the larger the DEf and the imprecision of the estimate will be. This will in turn make the demonstration that a particular objective has been reached more difficult. There is not one single strategy that totally controls for this effect, but some orientations are suggested for improving the sampling strategy *within* clusters:

Empirically, the clearest benefit in terms of reducing the DEf has come from selecting the next third, or next fifth from the first selected central household (instead of sending surveyors from that central household on to the next one directly). ⁵

The random selection of households within a village can be initiated from different points: Selection of an initial household from the periphery of the village in addition to a central starting point has been suggested. The results (in terms of gain in precision) are not perfectly satisfactory and will depend on the size and type of

the village.

Homogeneity will be decreased if the randomization procedure is started not from one central household but from four households, each central to a quadrant of the village.

Stratification, which we will discuss in the next section, because it pools respondents from each subgroup in each cluster, also decreases the homogeneity within clusters.

Overall, the DEf will remain an issue in cluster surveys, and its importance will vary with the clustering tendencies of each indicator and with the settings. Program managers and evaluators should develop a sense of its importance and consider possible strategies for diversifying the sample within clusters. Unfortunately they will not be able to validate the pertinence of their strategies without the use of statistical software (such as EPI-INFO), capable of calculating the DEf.

1.4. Options for managing sampling questions within sub-groups

A number of questions in the KPC survey refer not to the entire population surveyed, but to specific sub-strata in the sample (e.g. children 0-4 months or 0-6 months for exclusive breastfeeding). The sample size in any sub-stratum is obviously going to be lower than the total sample size. This will reduce the precision of the estimate within this subgroup and reduce the power of our tests. In other words, we may not be able to answer whether the objective for exclusive breastfeeding has been reached (the targeted percentage of exclusive breastfeeding mothers will be within the confidence interval of our estimate).

When it is of particular importance for the program to assess a situation with precision in that particular stratum, different alternatives are possible (table 3):

stratification,

increasing the overall sample size,

over-sampling in the specific age-group,

parallel sampling.

Table 3: options for increasing sample size in sub-groups

Method	Description	Advantages	Constraints
Stratifying	The population is divided in as many strata as specific sub-groups of interest (sex, age-groups). The sample is predetermined to include a defined number in each stratum, with the same contribution from each cluster.	The number of respondents in each stratum is known beforehand, the level of precision of the survey is improved (by decreasing the DEf), and no group is under-represented. Simple and straightforward.	Adds complexity to the sampling strategy.
Increasing the entire sample size	<p>The entire sample size is increased, and it is expected that the sample size of each sub-group of interest will also be increased proportionally.</p> <p>The number of clusters can be increased, or the number of respondents within clusters can be increased.</p>	<p>Precision will be improved by an increase in the number of clusters, as opposed to an increase in cluster size.</p>	<p>Increasing the number of clusters is costly.</p> <p>Increasing the size of the clusters will increase the DEf.</p> <p>Leaves to chance the selection of a sufficient number of children in a given age group.</p>
Over-sampling in the age-group of interest.	To have a precise coverage rate for immunizations among 12-23 months old children, the surveyors will be instructed to survey 10 children in each cluster, as they would normally, but then to interview additional mothers, exclusively about the EPI questions, until 7 children in total have been surveyed in the 12-23 months old age group. (In this case, this strategy would achieve a sample size of 210 children for the EPI questions). In most instances, this will only require adding two more children in the cluster. These two children will only be included in the analysis of the EPI questions.	Since these additional two are surveyed exclusively on the questions of immunization coverage, the added time (and cost) is relatively limited.	
Parallel sampling	A specific survey is administered for two different age groups. The sample size is calculated for each one separately, and two different questionnaires are prepared. The mothers are sampled from the same clusters and the same households and the two surveys are conducted through the same surveyors, using the same logistics. For each group, the desired level of precision is chosen and determines the size of the cluster for the age group.	<p>Cost-efficient use of logistical resources to obtain a predetermined level of precision in two different groups.</p> <p>Similar to stratification.</p>	Requires two questionnaires.

Over-sampling and parallel sampling represent cost-efficient approaches to reaching a satisfactory level of precision for important programmatic questions. The increases in cost and time of data collection are limited, and the data analysis process is not excessively complex. Once more, the need to provide information for management decision should lead the choice of the questions and that of the sampling strategy.

2. Management information at the local level and Lot Quality Assurance Sampling (LQAS)

2.1. A comparison of LQAS and cluster-sampling

One of the limitations of cluster-sampling in surveys is in the provision of management information at the local level. To illustrate this issue, let us imagine a PVO program overseeing the work of 10 local organizations, each responsible for one area of the region of intervention, with the objective of increasing measles' immunization coverage (figure 10). Supposing that the measles' immunization rate for the region has been established at 50% \pm 10% through a 30-cluster KPC survey.

Figure 10: Management information at the local level through cluster surveys – illustration.

As illustrated in our example, clusters may not correspond to a specific area of intervention. Clusters are not a representative random sample from each area for which we need estimates. Additionally, the size of the sample within a cluster is small and does not allow us to make estimates. The cluster sampling method helps the manager determine the scope of a problem for the entire program, but does not provide information about which local areas (which supervisors, or which sub-contracting agency) are performing better, and which require additional investments in training or other forms of support. As resources are limited, the program would be interested to know which areas are causing the overall low coverage rate, to take corrective actions, and which are achieving a much higher rate, in order to analyze the reason for their success.

LQAS (Lot Quality Assurance Sampling) provides an answer for these questions, and also gives an overall regional coverage rate estimate with a superior or equal precision

level. The reader is referred to other publications for an in-depth explanation of LQAS, its guiding principles and its development. A brief presentation of LQAS is given in Box 4. We will focus on illustrating what information can be provided through LQAS.

Keeping within the same illustration, let us now consider that local area supervisors conducted the KPC survey after having received a two-day training about LQAS (figure 11).

Figure 11: Use of LQAS in KPC surveys

LQAS can provide information at the local level with a small sample because it seeks to answer whether a given coverage is above (or below) a particular threshold judged satisfactory (or unsatisfactory), but does not provide an estimate of coverage at this local level. In other terms, managers can use LQAS not to estimate the coverage within a given area but to answer whether the area is a “high performance” or a “low performance” area with respect to measles’ coverage.¹²

Table 4 offers elements of comparison of LQAS and cluster sampling. LQAS’s strength is in its ability to provide decision rules usable at the peripheral level, while cluster sampling offers a rapid and simple method to assess the situation at a regional or district level.

In terms of cost, the principal advantage of LQAS is the ability to decentralize data collection at the level of the unit of study of interest (for example the area under the responsibility of a supervisor). When this data collection is conducted through local supervision and monitoring processes, LQAS has been used to survey an entire region at a lower cost than cluster sampling. When used to obtain an overall regional coverage figure, or conduct a full KPC survey with a central team of surveyors, it has shown to take more time and to cost more than cluster-sampling.¹³

Box 4: Steps in applying LQAS

LQAS can be taught relatively quickly to field supervisors, in usually two days. Its implementation follows a series of steps (A training manual for using LQAS to manage decentralized health programs: a user handbook.. Valadez J.J., 1998):

Define the service to be assessed.

Identify the unit of interest: a supervisor area, a district, a health worker?

Define the higher and lower thresholds of performance. A difference of 30% (25% to 35%) is recommended between the two thresholds, to maximize the efficiencies of LQAS. These thresholds are based on a management decision, and prior information about the expected performance of health workers.

1. Determine the level of acceptable error (risk of misclassification).
2. From a table determine the sample size and decision rule (number of errors accepted before an area is classified as performing “below expectations”).

For each area, the number of errors observed (non immunized children, service delivered without respecting the defined standards of care, etc.) will determine reliably if the area is performing above or below expectations.

Depending on the respective demographic weight of each area, an aggregated confidence interval for the regional estimate can be computed.

2.2. A brief review of LQAS applications in the literature

LQAS's main applications in child survival are service adequacy, coverage, and quality.^{12,14} Reported applications present LQAS as a practical, relatively low-cost field method that is increasingly being applied in health programs.¹⁵ The method has been used to assess immunization coverage, women's health issues such as family planning and antenatal care, use of ORT, disease incidence, and evaluation of health worker performance, in urban zones, rural areas, or on a national scale, in over 32 countries and the five continents. Lots have been defined as health center catchment areas, townships, villages, districts or zones in a city or within a province. Physicians or individual community workers have also been considered as lots in at least five surveys. Total sample sizes are reported from 70 to 25,230

Management information at the peripheral level

The first strength of LQAS is in providing information at the local level, which the supervisors can use immediately to assess either the performance of health workers, or local levels of coverage, knowledge and practices (whether above or below a certain prevalence). Conducting the baseline KPC survey, rather than a one-time exercise, becomes basis for establishing a health information system for the project. Comparisons can be made between service areas. Data gathering can be decentralized at the level of the supervisors, and be used rapidly through a decision process defined a priori. The literature offers examples of how the method can be integrated into supervision systems.¹⁶⁻¹⁹

Regional level coverage estimates by aggregating local data

The estimates at the local level are based on too small a sample to provide a useful level of precision, but they can be aggregated (and weighed according to the respective size of each zone) to provide a reliable regional estimate of coverage (or performance, whichever has been under investigation).

In summary:

LQAS is a sampling strategy designed to guide management decisions, and answer with confidence whether specific areas perform below or above a determined threshold.

The risk of error— as in all statistical procedures — is determined beforehand, by the manager / evaluator.

Because it does not attempt to provide precise estimates at the local level, but only to answer a Yes/No question, LQAS allows working with small samples.

In fact, sample sizes of 19 (19 households, 19 health workers, 19 activities performed by a health worker) offer a high level of precision for decision-making. When all areas under study perform above expectation, the LQAS decision rule loses its value as a way to differentiate high and low performance areas.

Let us imagine that the expected level of performance for measles coverage in all

our supervision-areas is 70%. A 19:9 decision rule (meaning that nine non-immunized children would be “accepted” in each sample of 19 children randomly selected from one supervision-area) will allow us to discriminate low from high coverage areas with a high specificity. But if all supervision-areas are achieving 70% coverage, all samples will have no more than nine non-immunized children, and no management decision will be made from the results. Having chosen a sample size of 19 it will, however, be possible to change our decision rule to 19:7 for example, and thus differentiate supervision areas based on an expected coverage of 80%.

A sample size of 19 allows the manager to retrospectively change the thresholds of acceptable performance, change the number of acceptable errors accordingly, and conduct a test on more stringent standards, in order to define which areas require additional management support.

Data collected at the local level can be weighed and aggregated to provide a reliable regional-level estimate. Because LQAS is very similar to a direct random sampling strategy, it does not have a design effect, and can provide a more precise estimate than cluster sampling methods (see example 7, in section 3.).

Table 4: a comparison of LQAS and cluster sampling

	LQAS	CLUSTER SAMPLE
Used to	Identifying high and low performance local areas in service age, and quality at the peripheral level assessing service adequacy and coverage at regional or national	assessing service adequacy and coverage at regional or national
Sample size	N = 15 to 28 per local area; (19 recommended as optimal in CS) Total sample by aggregation = 19 x m areas.	N = 30 * 7 = 210 for EPI surveys N = 30 x 10 = 300 for KPC surveys
assumptions	All children of study area under the care of the observed health	All children of study area under the care of the observed health
Advantages	Provides information at unit level, rapidly and locally Can assess CI for coverage at regional level* Sensitivity high for detecting low performance (high NPV) performance) Specificity increases with repeated measures at a local level (for h supervision) Data collection can be integrated with supervision system	Uniform level of coverage Rapid, uncomplicated, well known
Limits and constraints	Cannot draw CI for local area coverage (N too small) but simply estimation about extreme performance Sample size, is function of size of problem and cost of error Sampling frame: expensive and time consuming Specificity can be low for good performance	Does not identify local areas or units of low coverage or Requires 2 nd stage to identify problem units Can be biased by the selection of low or high performance clusters Large confidence interval of estimate (Design Effect)
Cost	More expensive and longer (3 times) than EPI-cluster for entire	Cheaper than random sample (might be cheaper than LQAS when
Threats to validity	Can be reduced by decentralizing data collection (data collection than conducting a cluster survey in this case)	Identical

3. *Qualitative questions in the KPC survey*

Child survival programs are frequently faced with questions about the cultural factors influencing the success or constraining their efforts to promote appropriate behaviors at the household level. They may also need information about ‘sensitive issues’, such as factors of risks in sexual behaviors, which are unlikely to be answered through survey questions. The KPC survey actually advocates that such issues be addressed through qualitative methods. Focus groups are the most commonly used method, but the potential benefits of key-informant interviews and other anthropological approaches, from which Participatory Learning and Action (PLA) has also derived, should not be neglected.

There simply is not a shortcut that would allow managers and evaluators to assess cultural and sensitive behavioral issues through survey questions. The investment in qualitative research needs to be thought about carefully, because of the limitations in the ability to generalize the data. But managers should also feel confident that validity is not only defined by the use of statistical procedures. Sampling for qualitative research has its own set of guiding principles and usually relies on non-probabilistic approaches.

It is beyond the scope of this paper to describe qualitative methods and the sampling approaches they require, and the reader is encouraged to refer to the relevant literature for a more in-depth discussion.

4. *Annotated sampling examples*

In this section we will present seven sampling examples to illustrate some of the key principles discussed throughout the report.

As a starting point, we will consider a PVO implementing a child survival program in a district covering 10 sub-districts. In accordance with the policies of the Ministry of Health, decision-making is being decentralized as much as possible to the sub-district level. The sub-districts are expected to raise most of their operating budget through user fees.

We will now present potential objectives for monitoring and evaluation, and demonstrate the consequences of these objectives in terms of sampling and sample size.

We will use measles vaccination coverage in children 12-23 months of age as an example of an indicator we are monitoring:

Example 1: sample size determination for baseline survey.

Example 2: assessing achievement of objectives in the entire district.

Example 3: analyses based on a final coverage estimate.

Example 4: calculating a sample size for the comparison between two groups

Example 5: assessing the significance of an observed change.

Example 6: an operations research study to compare two intervention approaches to improve immunization coverage.

Example 7: identification of sub-districts performing below standards and aggregated regional coverage estimate, using LQAS.

Example 1: Sample size determination for baseline survey.

From the information available in other regions, we expect measles vaccination coverage to be around 40%, but we would like to assess the coverage in our region of intervention within 10% of our estimate.

If we used simple random sampling to estimate this coverage, the appropriate formula for the sample size would be:

$$N = Z_{\alpha}^2 pq / d^2 \quad (1)^{(*)}$$

(*) Refer to appropriate reference for a discussion of the formula.

We can set the formula values as follows:

$Z_{\alpha} = 1.96$ corresponding to a confidence level of 95%

$p = 0.4$ (our expected coverage)

$q = 1 - 0.4 = 0.6$

$d = \text{accuracy desired} = 10\% = 0.10$

We obtain:

$$N = (1.96)^2 \times 0.4 \times 0.6 / (0.10)^2 = 92$$

For reasons of economy, time and logistics, we have decided to use a 30-cluster sampling method to conduct our survey. As we have seen, this introduces a design effect (DEf) in the precision of our estimate. A measles vaccination cluster survey in a neighboring region obtained a DEf of 1.8, slightly lower than the value of 2.0 usually used to calculate cluster-survey sample size.

We can now correct the sample size needed in our cluster survey (N_c) to achieve the same level of precision of 10% by using the formula:

$$N_c = N * DEf \quad (2)$$

In this case,

$$N_c = 92 \times 1.8 = 166$$

This is, of course, the sample size that would be needed in the age group concerned by measles vaccination. As the KPC survey targets children 0 to 23 months of age and not only 12 to 23 months, we need to obtain a total sample size large enough to include 166 children in the 12 to 23 months of age sub-group. If we estimate (from available demographic data) that 45% of the sample of children 0 to 23 months of age will be in the target age range for this indicator of 12-23 months of age, different options are available to ensure this result:

We can increase the total sample size proportionally to our need for 166 children aged 12-23 months.

$$N_t = 166 \times (100/45) = 369 \quad (3)$$

Where N_t is the total sample size.

In this case, we expect to have an appropriate sample size for our immunization coverage question.

As described in table 3 of section 3, we could also specifically oversample children aged 12-23 months so that 166 are sampled.

This approach is quite simple and cost-effective.

If it appeared important to ask one series of question for children 0-11 months and another for 12-23 months, a parallel sampling strategy might be used. In this case, two different sample sizes should be calculated and we would use $N_c=166$ for the immunization coverage question.

If we could decrease the homogeneity within each cluster, either by stratifying by age group, or by improving the recruitment process (selection of each third or fifth household after the first one, initiating the randomization from different quadrants of the villages/clusters) we would decrease the DEF of our survey. this approach is in fact

feasible can only be established by experimentation in similar settings, about similar questions, and analysis with a computer software such as EPINFO.

Assuming we expect to have a lower DEf, for example 1.2 instead of 1.8, we would then need a sample size of:

$$N_c = 92 \times 1.2 = 111 \quad (\text{see (2)})$$

$$N_t = (92 \times 1.2) \times (100/45) = 246 \quad (\text{see (3)})$$

Example 2: Assessing achievement of objectives in the entire district.

We now have to determine the sample size for the final survey in order to assess whether a target measles vaccination coverage of 70% has been reached.

With the CSAMPLE program in EPI-INFO for the baseline survey (with 30 clusters where every 3rd household was selected), we found a DEf of 1.5. We should assume the same DEf for our final survey if we follow the same method.

If we used simple random sampling with

$Z_{\alpha}=1.96$ corresponding to a confidence level of 95%

$p = 0.7$ (our target coverage)

$q = 1-0.7 = 0.3$

$d = \text{accuracy desired} = 10\% = 0.10$

We would need:

$$N = (1.96)^2 \times 0.7 \times 0.3 / (0.10)^2 = 84$$

Using a 30 cluster sample, we obtain:

$$N_c = 84 \times 1.5 = 126$$

Using the same logic as for the baseline we would need a total sample N_t :

$$N_t = 126 \times (100/45) = 280$$

Or we could also simply over-sample in the 12-23 months old age group (see example 1).

Example 3: Analyses based on a final coverage estimate.

Let us now assume that 137 children aged 12-23 months were in our final survey. If 78 of them (or 56.9%) have been vaccinated against measles, we can use the CSAMPLE program in EPI-INFO, to obtain a 'correct' 95% confidence interval (as opposed to calculating an 'incorrect' confidence interval by ignoring the DEf introduced by the cluster design). We obtain a DEf of 1.06 and a 95% confidence interval of our estimate between 48.4% and 65.5%.

We conclude that we have failed to reach our target coverage of 70%.

If we only 'recruited' 97 children aged 12-23 months in our final survey and 56 of them (or 57.7 %) have been vaccinated against measles. With a larger DEf of 1.70, we would obtain a 95% confidence interval of our estimate between 44.7% and 70.7%.

We cannot conclude (statistically) that we have failed to reach our target coverage of 70%, at the 95% confidence level. But our best estimate is that we are 12% below our objective. Accepting a smaller confidence level (90% for example) we could, however, probably reject having reached our target, since its value is close to the margins of our 95% confidence interval. But it would have been more satisfactory to increase the sample size in the age-group and to decrease the DEf, in order to be able to answer conclusively at the traditional 95% confidence level.

With 59 children immunized out of 65 (90.8 %), and a DEf of 1.38, we would obtain a 95% confidence interval of our estimate between 82.5% and 99.0%.

We conclude that we have reached our target coverage of 70% and are even above an 80% target, with a 95% confidence.

Inversely, in spite of a high estimate on a larger sample (79.4% or 104 children out of 131), a large DEf (e.g. 2.16) related to a high level of clustering of immunization, would yield a confidence interval between 69% and 89.6%, and would not allow us to conclude statistically that our estimate is statistically significantly superior to our preset objective. As a manager, we would report that our best estimate is that the

region of intervention has reached 79% coverage and that there is no statistical evidence against the program having reached its objective.

Example 4: Calculating a sample size for the purpose of a comparison between two phases or two groups.

Let us continue our example, using the 30-cluster method to conduct our surveys. Let us assume that a survey in a neighboring region using a 30-cluster design obtained a DEF of 1.8.

One formula given for calculating the two sample sizes is given by:

$$N_1 = N_2 = \frac{[Z_{\alpha/2} \sqrt{2pq}] + Z_{\beta} \sqrt{p_1q_1 + p_2q_2}]^2}{(p_1 - p_2)^2} \quad (4)$$

Where:

N_1 = baseline sample size

N_2 = final evaluation sample size

$Z_{\alpha/2}$ is the Z value corresponding to the chosen level of risk α . ($Z_{\alpha/2}$ should be used in two-sided tests, and Z_{α} should be used in one-sided tests.)

Z_{β} is the Z value corresponding to the chosen level of risk β (it directly relates to the 'power' of the test as power = 1 - β); (Z_{β} = 1.28 for a power of .9)

p_1 is the expected coverage at baseline

$q_1 = 1 - p_1$

p_2 is the expected final coverage

$q_2 = 1 - p_2$

$p = (N_1 p_1 + N_2 p_2) / (N_1 + N_2)$

$q = 1 - p$

In fact more precise statistical software use a correction of formula (4), as follows:

$$N_1' = N_2' = N_1 \times \left[1 + \frac{(1 + 4(p_1 - p_2))}{4(p_1 - p_2)^2} \right]^2 \quad (5)$$

If the expected coverage at baseline is 40%, and we want to be able to demonstrate an increase of 20 percentage-points in the final evaluation (meaning that we

want to be able to demonstrate an increase from 40% to 60%), the sample size for each survey would be:

$$N_1 = N_2 = \frac{\{1.96 \sqrt{2(.5)(.5)} + 1.28 \sqrt{[(.4)(.6) + (.6)(.4)]}\}^2}{(.4-.6)^2}$$

$$N_1 = N_2 = 129$$

A simplified formula is available, and would yield a similar result:

$$N_1 = N_2 = [Z_{\alpha/2} + Z_{\beta}]^2 [2pq] / (p_1 - p_2)^2 = [1.96 + 1.28]^2 [2(.5)(.5)] / (.2)^2 = 131$$

[Where p is the estimate sample proportion, and can be set at .5 if we make no assumption about the baseline and final coverage rates (p x q is maximum for p = q = .5).]

Statistical software, using formula (5) would yield $N_1' = N_2' = 140$.

140 children would be needed in the 12-23 months old group of interest for the baseline and final survey, if the samples were drawn by a simple random procedure. The cluster design forces us to correct the sample size in order to maintain the level of precision.

$$N_{1c} = N_{2c} = 140 \times 1.8 = 252$$

If we simply increased the total sample size in order to achieve 232 children in the 12-23 months age group, by the same process as in the preceding example, we would need a sample size for baseline and final evaluation of:

$$N_{1t} = N_{2t} = 252 \times (100 / 45) = 560 \text{ children.}$$

Example 5: Assessing the significance of an observed change

Let us now consider a situation where the baseline and final samples were chosen as:

$$N_1 = N_2 = 166$$

Our coverage rate estimates are 40% and 60% respectively at baseline and final, and we would like to assess whether this increase reflects a true change in the population of intervention. This question is similar to asking what the significance of the observed change is.

Our best estimate of the difference between the two proportions is: $(.6) - (.4) = .2$

If we ignored the DEF, we could construct a 95% confidence interval for the difference between the two proportions, with the following formula:

$$95\% \text{ CI for } (p_1 - p_2) = (p_1 - p_2) \pm Z_{\alpha} \times \sqrt{[(p_1 q_1) / N_1] + [(p_2 q_2) / N_2]} \quad (6)$$

In this case, we would obtain:

$$95\% \text{ CI for } (p_1 - p_2) = 0.2 \pm 1.96 \times \sqrt{[(.4)(.6) / 166] + [(.6)(.4) / 166]}$$

$$95\% \text{ CI for } (p_1 - p_2) = 0.2 \pm 0.105$$

$$\text{Lower } 95\% \text{ CI for } (p_1 - p_2) = 0.095$$

$$\text{Upper } 95\% \text{ CI for } (p_1 - p_2) = 0.305$$

The 95% CI (0.059 to 0.341) does not include zero, so we are 95% confident that a true increase of coverage rate has taken place. Our best estimate for this increase is 20%, and the 95% confidence interval is 9.5% to 30.5%.

NOTA:

(a) Alternatively, a Z-test can be conducted to test whether the two proportions are equal to one another. This is equivalent to constructing a 95% CI and observing whether

it includes zero or not.

$$Z = (p_1 - p_2) / \sqrt{[(pq) / N_1] + [(pq) / N_2]}$$

$$\text{with } p = (N_1 p_1 + N_2 p_2) / (N_1 + N_2)$$

Z can then be compared to a critical Z (e.g. 1.960 for a 5% significance level with one degree of freedom), which can be found in statistical tables.

(b) A more precise formula for the 95% CI is actually:

$$95\% \text{ CI for } (p_1 - p_2) = (p_1 - p_2) \pm Z_{\alpha} \times \sqrt{[(pq) / N_1] + [(pq) / N_2]}$$

$$\text{with } p = (N_1 p_1 + N_2 p_2) / (N_1 + N_2).$$

In reality, our sample did not come from a direct sampling method, and using the appropriate statistical software, the confidence interval would be corrected by a factor of the Def., that we will simply call C for this illustration. (For more precision about this correction factor, see Donner and Klar^{20,21}).

$$\text{True 95\% CI for } (p_1 - p_2) = (p_1 - p_2) \pm Z_{\alpha} \times \sqrt{[C \times (p_1 q_1) / N_1] + [C \times (p_2 q_2) / N_2]}$$

Depending on the Def, we would obtain a possibly much large CI, such as:

$$\text{true 95\% CI for } (p_1 - p_2) = 0.2 \pm 0.205$$

$$\text{Lower 95\% CI for } (p_1 - p_2) = -0.005$$

$$\text{Upper 95\% CI for } (p_1 - p_2) = 0.405$$

The true 95% CI (-0.005 to 0.405) includes zero so, although our best estimate for the difference of coverage between the two phases is 20%, we cannot conclude that it is statistically significantly different from 0.

Example 6: An operations research study to compare two intervention approaches to improve immunization coverage.

So far we have been considering sampling and evaluation designs for monitoring and evaluation of routine activities of a PVO child survival project. PVOs are increasingly becoming involved in operations research. The standards of proof required for a research study are higher, and compel us to use a formal study design.

We will now consider that the program wants to evaluate whether a new approach for carrying out routine immunization is significantly better than the approach currently used by the national immunization program. This change will have consequences in terms of costs, management and funding for the program. The government and other NGOs operating in the country will only be interested in adopting it if there is proof that it is superior to the current approach. The program manager therefore wants to be able to establish that the new approach was the “cause” of anticipated improvements in coverage, and that the improvements cannot be attributed to other causes.

The analysis is complicated in this case by the issue of “unit of analysis”. Since the sub-district was the unit of randomization and the level of intervention, the sub-district needs to be the unit of analysis. And in fact our sample size is five intervention plus five comparison sub-districts for a total of ten. Although we are administering KPC surveys to individual mothers about their individual children, our sample size is not the total number of children aged 12-23 months in the survey because individual children were not randomized to receive one immunization approach or another: All children living in one sub-district receive the same approach.²

If we were to not randomize which sub-districts would be the first to receive the new immunization approach, our data might be difficult to interpret. For example, if we were to implement the new approach in five sub-districts where the PVO had already been working in a previous food security project, we would not know if the better results achieved in the sub-districts receiving the new approach were due to the approach itself, or the trust and infrastructure established between the PVO and the communities during

the previous project.

We need an estimate for each sub-district with a precision of $\pm 10\%$. If we assume coverage of .5 for our calculation, formula (1) in example 1 gives us the desired sample size per sub-district:

$$N = 96$$

We need then a total sample size of

$$N_t = 96 \times 10 = 960 \text{ children.}$$

The Wilcoxon's Rank Sum Test is a non-parametric test, which allows us to compare results aggregated to the sub-district level. Each sub-district is given a rank (from one to ten) in descending order depending on its aggregate coverage rate. Tied ranks are allotted the mid-rank of the group. The ranks of the sub-districts are added for each group (intervention and control or comparison in this case) (table 5).

Table 5: Illustration of the Wilcoxon Rank Sum Test

		Pre-intervention				Post-intervention			
		Intervention		Control		Intervention		Control	
		Sub-districts		Sub-districts		Sub-districts		Sub-districts	
		coverage	rank	coverage	rank	coverage	rank	coverage	rank
		25%	1.5	25%,	1.5			26%,	1
								38%	2
		38%	4	34%	3	40%	3		
		41%	5					47%	4
		45%	6.5					50%	5
				45%	6.5	51%	6		
				49%	8			52%	7
		53%	10	50%	9	53%	8		
						54%	9		
						58%	10		
Sum of ranks		Ti =27.0		Tc=28.0		Ti2=36		Tc2=19	
Using the table of critical values of the rank sum test, we have the probability (p) that the smallest sum of rank is less of equal to Wo corresponding to Ho, where there is no difference between the two sums.									
Test difference (p-value)		p = 0.50 [ns]				p = 0.047			
						The difference is significant.			

Tables of critical values are available to determine whether a difference in the sum of the ranks is significant or not.²² (The process for using this test is also described in Smith and Morrow,²) It is expected that the sum of the ranks will not be statistically different at baseline, if the random allocation process has been respected. The computation is repeated for the post-intervention coverage rates, and the comparison is repeated, allowing us to conclude whether the observed changes were significant. The Wilcoxon's Rank Sum Test does not describe the amplitude of the change (in fact it does not consider the coverage rates at all, except for the purpose of ranking districts).

Example 7: identification of sub-districts performing below standards and aggregated regional coverage estimate, using LQAS.

Let us now consider that we want to be able to compare the performance of the ten sub-districts in order to select the sub-district supervisors with the most pressing need for increased support, including a refresher course. In this situation, LQAS will provide us with a rapid and cost-efficient way of gathering such information and it will also allow us to measure a global regional coverage rate.

LQAS requires following a step-by-step approach in defining thresholds of performance, acceptable risks of error, from which decision rules are made. Considering the small sample size with which LQAS operates, it operates with confidence for decision rules based on a substantial difference in performance level. For this reason, two thresholds of performance are determined: a higher threshold (the expected level of performance we want all sub-districts to reach), and a lower threshold, which will define sub-districts in need of immediate attention.

For our example, the program management team has decided that 80% measles coverage is the expected level of performance for all ten sub-districts. And immediate action will be taken to improve the situation in all sub-districts where coverage is below 50%. The two thresholds have now been determined.

Following the LQAS approach, the management team coached by the evaluator choose the risk of misclassification of sub-districts that they are willing to live with, and accordingly come to a decision rule or 19:6. Nineteen households will be randomly selected by the supervisor of each sub-district for control of the measles' immunization status. If six or less than six children among the 19 fail to be immunized, the sub-district will be classified as "above expectation". Our confidence level that the correct decision will be made in this case is 93.2%. If seven or more children among the 19 are not immunized, the sub-district will be considered as "below expectation", with a 91.6% confidence level.

It will take about two days to train the supervisors in the method, and about the same amount of time for them to collect the data, and immediately obtain the status of their sub-district based on the decision rule.

The program manager will thus obtain extremely quickly a performance map for all ten sub-districts under his/her responsibility.

If now, a coverage rate must be computed for the entire district from the sub-district level data, the information can be provided by a total sample size of:

$$N_t = 19 \times 10 = 190$$

The computation of the regional coverage estimate and its confidence interval requires us to use a weighing scheme based on the distribution of the population of children 12-23 months among the 10 sub-districts (table 6).

The regional coverage estimate (pr) is the sum of the weighted sub-district coverage estimates, where the weight (wt) is the ratio of the sub-district 12-23 months population over the region's 12-23 months population.

Although none of the sub-districts' estimates (p) are reliable, due to the small sample size, the overall regional coverage estimate of 0.68 is quite reliable. Its confidence interval can be calculated by the formula:

$$95\% \text{ CI regional coverage rate} = pr \pm 1.96 \times \sqrt{[wt^2 \times (pq) / n]}$$

From table 6 we have: $[wt^2 \times (pq) / n] = 0.0012$

and so:

$$95\% \text{ CI } (pr) = 0.68 \pm 1.96 \times \sqrt{0.0012} = 0.68 \pm 0.07$$

Our regional coverage estimate is 68%, with a 95% confidence interval between 61% and 74%.

Three sub-districts (C, D & I) require immediate attention as they are

significantly below expectations.

With a more stringent decision rule (19:5, based on thresholds of 90% and 50%), four sub-districts (B, E, F & J) would maintain a level of performance above expectations, and might provide valuable insights.

Table 6: Aggregate coverage estimate from 10 LQAS samples of size 19

District	12-23 mo	Weight	# children	Sample	Coverage	wt x p	$(wt)^2 \times (p)(q) / n$
A	4000	0.08	6	19	0.68	0.05	0.0001
B	8000	0.16	5	19	0.74	0.12	0.0003
C	3000	0.06	8	19	0.58	0.03	0.0000
D	7000	0.14	9	19	0.53	0.07	0.0003
E	4000	0.08	3	19	0.84	0.07	0.0000
F	6000	0.12	5	19	0.74	0.09	0.0001
G	3000	0.06	6	19	0.68	0.04	0.0000
H	6000	0.12	6	19	0.68	0.08	0.0002
I	5000	0.1	8	19	0.58	0.06	0.0001
J	4000	0.08	5	19	0.74	0.06	0.0001
Total	50000	1		190	pr =	0.68	0.0012

Conclusion

Information is a rare and essential commodity for managers of CS programs.

Information can serve many purposes:

- . Exploratory:
 - .What is the situation (in the broad sense of the term) faced by a new project?
- . Planning:
 - .What should the priorities of our program be?
 - .What should the program's objectives be?
- . Monitoring:
 - .What activities are effectively being implementing?
 - .With what level of quality are they being implemented?
- . Evaluation:
 - .What are the achievements of the program?
 - .Have objectives been reached?
 - .Have key indicators changed between baseline and final evaluation?
 - .What is the plausible responsibility (effectiveness) of the program in the improvement of key indicators?
- . Research:
 - .What is the probabilistically demonstrated responsibility (efficacy) of an intervention design vs. no intervention, all other factors being held constants?
 - .What is the respective benefit of an intervention design vs. another, all other factors being held

constants?

The systematic use of KPC surveys in the last 10 years has contributed substantially to improving diagnostics, decision-making, monitoring and evaluation in USAID-supported CS programs. The 30-cluster method has been the most widely used approach to data gathering at the population level, and LQAS is growing in favor as a useful way of bringing the analysis to the local operational level. Unsurprisingly, with time and experience, managers have increased their demand for information, in order to improve the management process, but also to report with rigor and method on the achievements of their programs. Monitoring for results has become a central question.

This report has tried to make more explicit the premises and limitations of 30-cluster KPC surveys, to present options for overcoming some of these limitations either by improvements on the cluster sampling method, or by the use of LQAS as an alternative sampling approach. Perhaps as importantly, we have tried to emphasize the conceptual differences between evaluation for research and evaluation for management. Each additional level of methodological complexity comes with an increasing cost to the program. And each level of information bears a specific importance for program management. There is, for these reasons, no single solution to the question of program evaluation for results.

Program managers need to be guided by important principles in choosing an evaluation design:

- . Sound management:
 - . Will the data gathered (purchased) lead to a decision?
 - . What level of precision is required for this information to be useful?
 - . Is the cost of the information appropriate for the scale of the intervention?
- . Accountability:
 - . Will the information inform appropriately donors and partners?
 - . What level of precision and certainty should be expected from the program evaluation?
- . Development:

- . Is evaluation still a capacity-building tool?
- . Are the methods used to inform program management sustainable in the context of the intervention?

The choice of methods and tools should take place within a greater understanding of the methodological issues presented in this paper and be based on these essential principles.

Approaches to monitoring for results

Evaluation objectives of KPC surveys

SECTION 3: Sampling options for KPC surveys

Conclusion

Resources on methodology and sampling issues for surveys

- 1. Conducting small-scale nutrition surveys – A field manual. Nutrition planning, assessment and evaluation service. Food policy and nutrition division. Food and Agriculture Organization of the United Nations. Rome, 1990; 186 pages.**

Purpose of the manual

This manual has two primary purposes: 1) to assist the nutritionist in deciding whether or not to conduct a project-specific nutrition survey; and 2) to provide the non-survey specialist with practical step-by-step guidance in conducting a survey. The manual addresses the important issue of when it is appropriate and when not to conduct a survey.

Although the manual mainly addresses the specific needs of a project-specific nutrition survey, it also provides basic principles that can be applied to other kinds of surveys.

Organization of the manual

The manual is organized in eight chapters, opening with an introduction:

Chapters 1 and 2 provide an introduction to the manual and to surveys. Included are basic guidelines on how to decide whether or not to do a project-specific nutrition survey and a detailed list of all the steps in the survey process. Chapters 3 and 4 focus on planning the survey and selecting the survey sample. Discussed are the preliminary planning, budgeting and organization of the survey work and the survey team. Also provided is a general introduction in non-technical terms to the basic principles of sampling theory and explains step by step how to draw a statistically representative sample. Chapter 5, "Choosing Survey Content," describes some general categories of information usually collected in nutrition surveys and suggests the best ways to collect that information. The sixth and seventh chapters address the tasks of writing the questionnaire and collecting the data. Reviewed are the practical needs of recruiting and training interviewers and supervising data collection. Chapter 8, "Processing and Analyzing the Data," discusses various ways to interpret and analyze survey results and provides some

simple formulas for testing the validity and significance of data. The final chapter (9), "Presenting and Using the Survey Results" explains the importance of clear, concise and timely reporting and provides guidelines for writing the final report and presenting the survey results to project planners.

Ordering Information

Source #1

Price: \$33.00, ISBN 02851

Sales and Marketing Group
Food and Agriculture Organization
Viale delle Terme di Caracalla
00100 Rome, Italy

Telephone (+ 39 06) 57055727

Fax (+ 39 06) 57053360

E-mail (for orders) Publications-sales@FAO.Org

E-mail (for information) Nutrition@FAO.Org

Internet Order Form

<http://www.fao.org/CATALOG/interact/order-e.htm>

Source #2

Can be downloaded from the Anthropometry Resource Center:

<http://www.odc.com/anthro/>

[select "Anthropometric Desk Reference," then click on the section, "Reference Books"]

- 1. Constructing Samples for Characterizing Household Food Security and for Monitoring and Evaluating Food Security Interventions: Theoretical Concerns and Practical Guidelines. Carletto C. International Food Policy Research Institute. 1999; 35 pages.**

Purpose of the guide

This guide discusses how random sampling techniques can economize on the costs of gathering information while increasing the likelihood that it will be both accurate and available in a timely fashion. It is organized in two parts. The first part of the guide provides a non-technical overview that begins with a discussion of why random samples are often favored over non-random samples and censuses for obtaining information on household characteristics such as food security. Provided is a step-by-step description of the process of constructing a random sample. This description is followed by an example that outlines how a random sample of farmers was obtained in order to assess the impact of two projects directed toward smallholders in Malawi. The second part of the guide

consists of four technical appendices that complement and expand upon the discussion found in the overview: 1) glossary; 2) calculating sample sizes; 3) using random number tables to obtain a sample; and 4) selecting clusters when they are of unequal size. These appendices are designed for individuals who have some familiarity with statistics.

Ordering Information

Single copies available free of charge from:
International Food Policy Research Institute
2033 K Street, NW
Washington, DC 20006, USA
Telephone 202-862-5600
Fax 202-467-4439
E-mail ifpri@cgiar.org
Internet <http://www.ifpri.org>
(Available in PDF format:
<http://www.iimi.org/ifpri/themes/mp18/pubs.htm>)

2. How to Sample in Surveys. Fink A. The Survey Kit TSK 6. Sage Publications, 1995; 73 pages.

Purpose of the book

This book is designed to provide guidelines for selecting and using appropriate sampling methods for surveys. It offers a simple presentation of the principles and types of sampling methods. The specific objectives of the book are to help the reader: 1) Distinguish between target populations and samples; 2) Choose the appropriate probability and non-probability sampling methods (many sampling techniques are addressed, including simple random sampling, stratified random sampling, systematic sampling, cluster sampling, convenience sampling, snowball sampling, quota sampling, and focus groups); 3) Understand the logic in estimating standard errors; 4) Understand the logic in sample size determinations; 5) Understand the sources of error in sampling; and 5) Calculate the response rate.

Ordering Information

Price: \$14.95, ISBN: 0-8039-5754-8

3.

In the USA

Sage Publications, Inc.
2455 Teller Road
Thousand Oaks, CA 91320

Telephone 805-499-0721
Fax 805-499-0871
E-mail info@sagepub.com
Internet <http://www.sagepub.com/>

3. In Europe

Sage Publications, Ltd.
6 Bonhill Street
London, EC2A 4PU, United Kingdom
Telephone: +44(0)171 374 0645
Fax +44(0) 171 374 8741
E-mail: market@sagepub.co.uk
Internet: <http://www.sagepub.com/>

3. Monitoring Immunization Services Using the Lot Quality Technique. WHO Global Programme for Vaccines and Immunization. World Health Organization, Geneva, 1996; 119 pages + answer sheets (1v.).

Purpose of the manual

This manual is designed to provide guidance to managers who want to use the Lot Quality (LQ) technique to monitor immunization services. It focuses on two uses of the LQ technique: Lot Quality Assessment and Lot Quality Coverage Survey. LQ Assessment is done to decide whether one or more health service units are meeting a specified standard of performance. LQ Coverage Survey is performed to measure immunization coverage, which is done by aggregating data from all health service units in the area being surveyed. Other than the data aggregation process (discussed in Section 4.3), the steps for carrying out the LQ technique are the same for either use and are described in this manual.

4. The manual is organized in five sections that provide step-by-step instructions for conducting a household-based evaluation using the LQ technique: 1) Plan the survey; 2) Prepare for the survey; 3) Conduct the survey; 4) Tabulate and analyze data; and 5) Take action. Seven different exercises are incorporated into the manual for practicing the components of the Lot Quality Technique. Answer sheets to the exercises are provided in the supplement to this document.

Exercise A is worksheet that takes the reader through the steps leading up to a final estimation of the total sample size for a Lot Quality study. In *Exercise B* the worksheet started in Exercise A is completed when the reader determines a lot sample size, low threshold, high threshold, and decision value. During *Exercise C* the reader is asked to select sampling points as s/he would as a district level supervisor conducting a

Lot Quality Assessment in health centers. In *Exercise D* the reader is asked to describe how s/he would select a household in which to collect data in a sampling point area that is selected in a community with no individual or household list and no count of households. *Exercise E, F & G.* In these three exercises, the forms for each lot are checked and completed. Additionally, the reader will practice the steps involved with completing summary forms, including a worksheet on the aggregation of Lot Data.

Ordering Information

Document numbers: Main - WHO/VRD/TRAM/96.01
Supplement - WHO/VRD/TRAM/96.01 SUPP.1
World Health Organization
Global Programme for Vaccines and Immunization
CH-1211 Geneva 27
Switzerland
Telephone +41 22 791 43 73
Fax +41 22 791 41 93
E-mail gpv@who.ch
Internet
<http://www.who.ch/programmes/gpv/genglish/avail/gpvcatalog/catalog1.htm>

4. NGO Networks for Health Detailed Monitoring and Evaluation Plan. Valadez JJ, Plan International. A publication of the NGO Networks for Health Project. 1999; 64 pages.

Purpose of the document

This document offers a presentation of and guidelines for the principles and implementation of monitoring and evaluation in child survival projects, with discussion of the KPC indicators, and guidelines for using LQAS in monitoring health projects. Specifically, the document has four objectives: 1) To describe the Networks' Project Monitoring and Evaluation (M&E) Plan, including procedures; 2) To present an approach that PVOs can use to carry out high quality service provision through high quality M&E at the country level, either alone or in a network; 3) To present illustrative M&E indicators that can be used as *core* indicators that the Networks Project will report to the Global Bureau, Center for Population, Health and Nutrition (G/PHN), and other *priority* indicators that the Networks Project Management Unit (NMU) would use to monitor the project. Additional indicators, which are as yet untested, are also considered to facilitate program management; and 4) To indicate how the project will interact with USAID Missions with respect to M&E.

Ordering Information

5. *NGO Networks for Health*
1620 I Street, NW
Suite 900
Washington, DC 20006
Telephone 202-955-0070
Fax 202-955-1105
E-mail info@ngonetworks.org
Internet <http://www.ngonetworks.org>

5. Sampling Guide. Magnani R. Food Security and Nutrition Monitoring (IMPACT) Project, for the U.S. Agency for International Development. 1997; 47 pages.

This guide belongs to a series called the Title II Generic Indicator Guides that are part of the USAID's support of the Cooperating Sponsors in developing monitoring and evaluation systems for use in Title II programs. (Appendix 1 of this guide lists the Generic Title II Indicators.) The objective of this series of guides is to provide the technical basis for the indicators and the recommended method for collecting, analyzing and reporting on the generic indicators.

Purpose of the guide

The *Sampling Guide* is designed to provide guidance on how to go about choosing samples of communities, households, and/or individuals for (sample) surveys. The aim is to select samples that can be combined with appropriate indicators and evaluation study designs to reach valid conclusions about the effectiveness of Title II programs. This guide emphasizes the use of probability sampling methods.

This guide was written for readers with a limited background in sampling. However, knowledge of basic statistics will be useful. Materials are presented step-by-step in the order likely to be followed in carrying out a Title II evaluation. Four principal phases are described: 1) defining the measurement objective of the study; 2) determining the sample size requirements; 3) selecting the sample; and 4) analyzing the data.

Ordering Information

Copies available free of charge from:

Source #1

Food and Nutrition Technical Assistance Project (FANta)
Academy for Educational Development
1825 Connecticut Avenue, NW
Washington, DC 20009-5721
Telephone 202-884 8000

Fax 202-884 8432
E-mail fanta@aed.org
Internet www.fantaproject.org

Source #2

Food Aid Management
300 I Street, NE, Suite 212
Washington, DC, 20002
Telephone 202-544 6972
Fax 202-544 7065
E-mail fam@foodaid.org
Internet www.foodaid.org

Source #3

Copies can be downloaded in PDF version (requires Acrobat Reader) or in WordPerfect (self-extracting file) from the FANta web site,
<http://www.fantaproject.org/pubs.htm>

6. Sample Size Determination in Health Studies: A Practical Manual, Lwanga SK, Lemshow S. World Health Organization, Geneva. 1991; 80 pages.

Purpose of the manual

This manual provides a brief guide to selecting sample sizes for the most common situations encountered in health studies. It is designed for health workers and managers who do not have detailed knowledge of statistical methodology. In particular it is geared toward individuals working at the local or district level. Provided are several situations in which minimum sample size must be determined, including studies to estimate population proportion, odds ratio, relative risk and disease incidence. For each situation addressed, an illustrative example is given. In addition to these examples, over fifty tables are provided to help the reader determine the proper sample size. This manual is designed to be used in "cookbook" fashion as a practical guide to making decisions on sample size once a proposed study and its objectives have been clearly defined.

Ordering Information

Price: Sw. fr. 16.-, (Sw. fr. 11.20 in developing countries), ISBN: 92 4 154405 8
World Health Organization, Distribution and Sales
CH-1211 Geneva 27
Switzerland
Telephone +41 22 791 24 76
Fax +41 22 791 48 57
E-mail (For orders) bookorders@who.ch
E-mail (For questions) publications@who.ch
Internet <http://www.who.org/>

USAID Development Experience Clearinghouse
1611 N Kent St Ste 200
Arlington VA 22209-2111, USA
Telephone 703-351-4006 (extension 106)
Fax 703-351-4039

E-Mail docorder@dec.cdie.org
Internet http://www.dec.org/partners/dexs_public/about.cfm

7. Survey Trainer's Guide for PVO Child Survival Project Rapid Knowledge, Practice and Coverage (KPC) Surveys. The Johns Hopkins University, School of Hygiene and Public Health, PVO Child Survival Support Program. January 1997.
(run a search for the title)
DOCID/Order No: PN-A-CE-202

Purpose of the guide

This resource consists of a survey trainer's guide and a binder of 22 appendices. The guide is primarily designed for training staff of community or district level child survival projects (or other primary health care programs) to plan, carry out, and analyze a Rapid KPC Survey. It can also be used as a reference tool for trained staff as they carry out additional Rapid KPC Surveys. It systematically presents the theoretical and practical elements of information needed to train staff, conduct KPC surveys, tabulate and analyze results and develop action plans based on findings. The 22 appendices are meant to complement the guide by offering a suggested format and phased scheduling for a survey coordinator and a core team to learn how to train participants in all phases needed to conduct a Rapid KPC Survey. Each appendix contains a basic checklist of actions required in order to complete all phases of training. Provided are all the necessary components and instructions needed to complete a baseline or follow-up survey used in PVO child survival projects. Included in the appendices are references, information and exercises on sampling, sample size calculation and determination of statistical reliability.

Ordering Information

Price: \$12.87 (paper) | \$2.50 (microfiche)

8. A Training Manual for Using LQAS to Manage Decentralized Health Programs: A User Handbook. Valadez JJ, Plan International, 1998; 35 pages +tables.

Purpose of the manual

This manual offers a systematic and step-by-step presentation of LQAS and its application to managing health systems. It is written for individuals working in Ministries of Health and NGOs. Provided are guidelines for assessing community-based programs, as well as programs organized from health centers or regional hospitals. The manual is organized in eight chapters that address: what information is needed by a health program manager to assess a program's effectiveness; how to use LQAS; how to present an LQAS result; how to aggregate LQA samples to calculate coverage proportions with a confidence interval; and how to assess the technical skills of a health worker. An appendix provides LQAS tables for sample sizes ranging from 10-30.

For a more detailed review of the principles of LQAS, its application and its field testing for the assessment of child survival programs, see *Assessing Child Survival Programs in Developing Countries: Testing Lot Quality Assurance Sampling*.¹²

Ordering Information

9. *NGO Networks for Health*
1620 I Street, NW
Suite 900
Washington, DC 20006
Telephone 202-955-0070
Fax 202-955-1105
E-mail info@ngonetworks.org
Internet <http://www.ngonetworks.org>

9. References

1. Habicht JP, Victora CJ. Evaluation designs for adequacy, plausibility and probability of public health programme performance and impact. *International Journal of Epidemiology* 1999; 28:10-18.
2. Smith P, Morrow RH. Methods of analysis. In: World Health Organization, editor. *Field Trials of Health Interventions in Developing Countries: A toolbox*. 2 ed. MACMILLAN, 1996:323-328.
3. Henderson RH, Sundaresan T. Cluster sampling to assess immunization coverage: a review of experience with a simplified sampling method. *Bull World Health Organ* 1982; 60:253-260.
4. Brogan D, Flagg EW, Deming M, Waldman R. Increasing the accuracy of the Expanded Programme on Immunization's cluster survey design. *Ann Epidemiol* 1994; 4:302-311.
5. Bennett S, Radalowicz A, Vella V, Tomkins A. A computer simulation of household sampling schemes for health surveys in developing countries. *Int J Epidemiol* 1994; 23:1282-1291.
6. Le TN, Verma VK. An analysis of sample designs and sampling errors of the demographic and health surveys. 1997; Macro International Inc., Calverton, MD, USA. 3: Demographic and Health Surveys; Analytical Reports.
7. Katz J. Sample-size implications for population-based cluster surveys of nutritional status. *Am J Clin Nutr* 1995; 61:155-160.
8. Katz J, Carey VJ, Zeger SL, Sommer A. Estimation of design effects and diarrhea clustering within households and villages. *Am J Epidemiol* 1993; 138:994-1006.
9. Katz J, Yoon SS, Brendel K, West KPJ. Sampling designs for xerophthalmia prevalence surveys. *Int J Epidemiol* 1997; 26:1041-1048.

10. Katz J, Zeger SL. Estimation of design effects in cluster surveys. *Ann Epidemiol* 1994; 4:295-301.
11. Yoon SS, Katz J, Brendel K, West KPJ. Efficiency of EPI cluster sampling for assessing diarrhoea and dysentery prevalence. *Bull World Health Organ* 1997; 75:417-426.
12. Valadez JJ. *Assessing Child Survival Programs in Developing Countries: Testing Lot Quality Assurance Sampling*. Harvard University Press, 1991.
13. Singh J, Jain DC, Sharma RS, Verghese T. Evaluation of immunization coverage by lot quality assurance sampling compared with 30-cluster sampling in a primary health centre in India. *Bull World Health Organ* 1996; 74:269-274.
14. Lemeshow S, Taber S. Lot Quality Assurance Sampling: single- and double-sampling plans. *World Health Stat Q* 1991; 44:115-132.
15. Robertson SE, Anker M, Roisin AJ, Macklai N, Engstrom K, LaForce FM. The lot quality technique: a global review of applications in the assessment of health services and disease surveillance. *World Health Stat Q* 1997; 50:199-209.
16. Valadez JJ, Weld L, Vargas WV. Monitoring community health workers' performance through lot quality-assurance sampling. Letter. *Am J Public Health* 1995; 85:1165-1166.
17. Valadez J. Assessing technical quality of service delivery. In: Valadez J, editor. *Assessing child survival programs in developing countries. Testing log quality assurance sampling*. 1991:129-144.
18. Valadez J, Vargas W. Supervision of primary health care in Costa Rica: time well spent? *Health Policy and Planning* 1990; 5:118-125.
19. Valadez JJ, Transgrud R. Assessing family planning service-delivery skills in Kenya. *Stud Fam Plann* 1997; 28:143-150.
20. Donner A, Klar N. Confidence interval construction for effect measures arising from

cluster randomization trials. J Clin Epidemiol 1993; 123-131.

21. Donner A, Klar N. Methods for comparing event rates in intervention studies when the unit of allocation is a cluster. American Journal of Epidemiology 1994; 140:279-289.
22. Pagano Marcello, Gauvreau Kimberlee. Principles of biostatistics. Duxbury Press, Belmont, California, 1993.